

Predicting customer purchase in an online retail business, a Data Mining approach.

A

Thesis Report

Submitted in

partial fulfillment of the requirement for the degree of

Bachelor of Technology in Computer Science.

by

Aniruddha Mazumdar

(10606009)



Department of computer Science & Engineering

NATIONAL INSTITUTE OF TECHNOLOGY ROURKELA

ROURKELA – 769008 (ORISSA)

May 2010



National Institute of Technology Rourkela
Rourkela-769008 (Orissa)

CERTIFICATE

This is to certify that the thesis entitled, “**Predicting customer purchase in an online retail business, a data mining approach**” submitted by Aniruddha Mazumdar in partial fulfillments for the requirements for the award of **Bachelor of Technology Degree in Computer Science Engineering**, National Institute of Technology, Rourkela is an authentic work carried out by him under my supervision. The matter embodied in the thesis has not been submitted elsewhere for a degree.

Date:

Prof. Bibhudatta Sahoo
Assistant Professor
Dept. of Computer Science Engineering
National Institute of Technology

Rourkela – 769008

ACKNOWLEDGEMENT

On the submission of my Thesis report, I would like to extend my gratitude & my sincere thanks to my supervisor Prof. B.D. Sahoo, Assistant Professor, Department of CSE, for his constant motivation and support during the course of my work in the last one year. I truly appreciate and value his esteemed guidance and encouragement from the beginning to the end of this thesis. I would like to thank all the teachers and staff who made my stay in Rourkela an unforgettable and rewarding experience. Finally, I am grateful to my parents and friends for their motivation and support throughout my career.

Aniruddha Mazumdar

CONTENTS

Abstract	6
List of Figures	7
1. Introduction	8
1.1 The need for Customer Behavior prediction and data mining.	9
1.2 Relevance of Data mining towards CRM.	10
1.3 Related work.	11
1.4 Problem model formulation.	12
1.5 Layout of thesis	13
2. The Customer Segmentation approach.	14
2.1 An Introduction to clustering and customer segmentation.	15
2.2 A background of the Vector Quantization based clustering algorithm.	15
2.3 The Structure and function of the VQ algorithm used for the study.	17
3. The Association rules based approach for customer purchase predictions.	19
3.1 A background of the Association rules based data mining approach.	20
3.2 A Description of the Association rules mining model proposed here.	21
4. Implementation and results.	23
4.1 Sample dataset and the transformation of data.	24
4.2 Choice of Programming language and environment.	28
4.3 Study of the VQ based clustering algorithm. Results found and discussion.	29
4.3.1 conclusions and inference from above data.	30

4.4	Study of the Association rule based customer behavior mining Approach, Results and discussions.	31
5.	Conclusions.	37
5.1	Conclusion	38
5.2	Future work	39
	References	40

Abstract

Identifying segments of customers and their behavioral patterns over different time intervals, is an important application for businesses, especially in case of the last tier of the online retail chain which is concerned with “electronic Business-to-Customer relationship” (B2C) . This is particularly important in dynamic and ever-changing markets, where customers are driven by ever changing market competition and demands. This could lead to the prediction of ‘churn’, or which customers are leaving the company’s loyalty. Also, the provision of customized service to the customers is vital for a company to establish long lasting and pleasant relationship with consumers. It has also been observed that keeping old customers generates more profit than attracting new ones. So, customer retention is a big factor too. So, there is always a trade-off between customer benefits and transaction costs, which has to be optimized by the managers.

The purpose of this thesis is to study, implement and analyze various Data-mining tools and techniques and then do an analysis of the sample / raw data to obtain a meaningful interpretation. Some of the data mining algorithms I have used, are a vector quantization based clustering algorithm, and then an ‘Apriori’ based Association rule mining algorithm. The first one is aimed at a meaningful segregation of the various customers based on their RFM values, while the latter algorithm tries to find out relationships and patterns among the purchases made by the customer, over several transactions.

LIST OF FIGURES

Figure 1:	The basic CRM cycle .	11
Figure 2:	The sample dataset for the Association rule mining approach(some entires)	25
Figure 3:	Sample dataset for the VQ based customer clustering approach.	27
Figure 4:	the initial input data for VQ approach.	29
Figure 5(a):	The output observed for $e = 0.002$ for the above VQ approach.	30
Figure 5(b):	The output in a graphical format	30
Figure 6:	The occurrences for 1- itemset.	31
Figure 7:	The final rules obtained after pruning with coverage = 0.2.	32
Figure 8:	The final rules obtained after pruning with coverage = 0.5.	33
Figure 9:	Rule set with coverage = 0.7.	34
Figure 9:	The snippet of rule set for 3-itemset obtained for coverage = 0.5	35

CHAPTER 1 : INTRODUCTION

1.1 The need for Customer Behavior prediction and data mining.

The emergence of the business-to-customer (B2C) markets has resulted in various studies on developing and improving customer retention and profit enhancement. This is mainly due to the retail business becoming increasingly competitive with costs being driven down by new and existing competitors. In general, consumer markets have several characteristics such as repeat-buying over the relevant time interval, a large number of customers, and a wealth of information detailing past customer purchases. In those markets, the goal of CRM is to identify a customer, understand and predict the customer-buying pattern, identify an appropriate offer, and deliver it in a personalized format directly to the customer.

One typical example of the CRM model corresponds to the case of an online retail shop which sells various products through internet and performs transaction directly with customers through the internet. An online retail shop defines a customer as a person who has already bought products or performed a transaction with the shop. The exponential growth of the Internet has led to a hoard of customer and market data to the market managers. The increased availability of individual consumer data presents the possibility of direct targeting of individual customers. That is, the abundance of customer information enables marketers to take advantage of individual-level purchase models for direct marketing and targeting decisions. But, such an enormous amount of data can be a huge clutter, and it can become cumbersome to draw meaningful conclusions from such raw data. This is where the utility of customer behavior prediction using Data mining techniques comes in.

The major customer values or characteristics that are used to measure purchase behavior of customers include **Recency, Frequency, and Monetary values (RFM)**. **RFM** measures provide information on what customers do. Recency tells how long it has been since each customer made the last purchase. Frequency tells how many times each customer has purchased

an item during certain intervals of time. Monetary tells how much each customer has spent in total. Monetary measures the total expenditure of the customer for a number of transactions over a period of time. These characteristics may be the most important in determining the likely profitability of a particular product or an individual customer, so they are used to segregate the list of customers into groups having different characteristics based on the RFM values.

1.2 Relevance of Data mining towards CRM.

Data mining techniques are the processes designed to identify and interpret data for the purpose of understanding and deducing actionable trends and designing strategies based on those trends [3]. Data mining techniques extract the raw data, and then transform them to get the transformed data, and then get meaningful patterns among the transformed data. As businesses evaluate their investments on marketing activities, they tend to focus on their data mining techniques and capability. How to learn more about customers and their inclination towards particular products, use that information to make appropriate choices to customers, and understand which marketing strategies can succeed in long term customer satisfaction and retention. Managers can understand their customer by evaluating customer behavior, customer segregation, customer profiles, loyalty (how long have they been associated with the company) and profitability (which products can be targeted to the particular customer so as to extract maximum profits). Data Mining helps managers to identify valuable patterns contained in raw data and their relations so as to help the major decisions.

The basic structure of CRM model lifecycle is shown in fig.1. The model can have two initiating pts. Firstly, the customer does some purchase and then the data is measured and evaluated. Afterwards, the company mines the evaluated data and then they can have an understanding of the patterns that the customer shows while purchasing. With the help of that

data, the organization can formulate its steps to maximize or optimize its business plans. Secondly, the organization takes some action for improving the customer's satisfaction by making a good informative offer, and then studies the actions taken by the customer. Then the actions of the customer are again evaluated and an understanding of the customer is achieved.

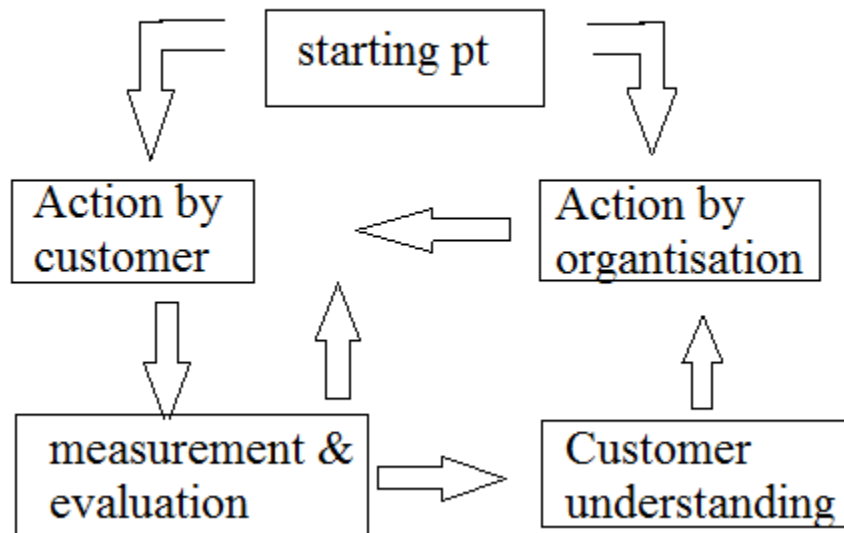


Fig1. The basic CRM cycle . [3]

1.3 Related work.

Customer behavior analysis is based on consumer buying behavior, with the customer playing three distinct roles of a user, payer and buyer. Relationship marketing is an important

aspect for customer purchase analysis as it has an importance in the research of the marketing through the re-affirmation of the importance of the customer or buyer. A greater importance is also placed on existing consumer retention, customer relationship management, personalization, customization and one-to-one marketing. Customer understanding is the heart of CRM. It is the basis for optimizing customer lifetime value, which in turn engulfs customer segmentation and actions to maximize customer conversion, retention, loyalty and profitability. Proper customer understanding and action ability lead to increased customer lifetime value. Improper customer understanding can lead to catastrophic actions. A customer can be a user, purchaser, influence-maker etc. Therefore the transaction data query may be have different types of inquiries, which include, suggestions, queries, requisitions, and reclamations. In the “Examination of Customer’s Inquiry” step, we can scrutinize which type of query has been placed by the customer and where it will be forwarded. In [3], Abdullah Al-Mudimigh, Farrukh Saleem and Zahid Ullah evaluate and analyze the customer buying pattern by using rule induction process on clustered data from the customer's database with reference to the customer query. In [5], Euiho Suh , Seungjae Lim , Hyunseok Hwang and Suyeon Kim lay their focus on studying anonymous customers and then they sequentially mine the data through data preprocessing and then extracting association rules from them.

1.4 Problem model formulation.

A typical online retail mart has thousands of transactions stored in its database. It handles hundreds, maybe thousands of customers per day. All these data transactions, also referred to as ‘orders’, need to be clustered according to some chosen parameters, and then a meaningful pattern or rules are to be inferred. For example, if a customer buys a certain product, is it necessary that he will buy another product related to that purchase, i.e., how to infer a rule

among 2 or more purchases of the customer. To incorporate a pattern to recognize a group of customers having similar purchase behavior.

This study focuses on the following aspects. Firstly, the study of the customer segmentation approach using a VQ algorithm, incorporated in the customer clustering approach. Secondly, the construction of a method of mining the data from raw datasets (a Microsoft access sample database in this case), and then transforming them into useful rules and inferences, for analyzing the customer's purchase behavior. And we shall later see that as the amount of raw data accumulates, and the level of rule association among various factors increases, the computational overhead for each of the algorithms also increases.

1.5 Layout of thesis.

The thesis is sub divided into 5 chapters. The 1st chapter is the introduction part which gives a background information on data mining techniques and the Customer Relationship model. The 2nd chapter deals with “customer segmentation” approach which encompasses the information on the VQ based algorithm and its basic working methodology. The 3rd chapter deals with the “customer purchase prediction” approach which explains the working of the Association rules mining approach. The 4th chapter “Implementation and results” shows the outputs of various algorithms mentioned in the previous chapters and the results according to various parameters. Finally the thesis is concluded by the 5th chapter “conclusions” by mentioning the finding the outcomes and laying the ground work for the future prospects of work in the same field.

CHAPTER 2: The Customer Segmentati- on approach.

2.1 An Introduction to clustering and customer segmentaion.

Customer segmentation is one of the most important area of knowledge-based marketing. In case of online retail stores, it is really a challenging task, as data bases are large and multidimensional. In the thesis we consider a clustering algorithm, which is based on a Vector Quantization based algorithm, and can be effectively used to automatically assign existing or new arriving customers into the respective clusters .

2.2 A background of the Vector Quantization based clustering algorithm.

It is an efficient algorithm designed by Linde, Buzo and Gray for the design of good block or vector quantizers with quite general distortion measurements is developed for use on either known probabilistic source descriptions or on a long training sequence of data, incorporated herein by reference, [1]. The algorithm involves no differentiation; hence it works well even when the distribution has discrete components, as is the case when a sample distribution obtained from a training sequence is used. As with the common variational techniques, the algorithm produces a quantizer meeting necessary but not sufficient conditions for optimality. Usually, however, at least local optimality is ensured in both approaches.

An N-level k-dimensional ‘quantizer’ is a mapping, q ; that assigns to each input vector, $x = (x_0, \dots, x_{k-1})$, a reproduction vector, $\hat{x} = q(x)$, drawn from a finite reproduction alphabet, $A = \{b_i; i = 1, \dots, N\}$ [1]. The level N describes the number of times the division of the codebook occurs. The quantizer is completely described by the reproduction alphabet (or codebook) A together with the partition, $S = \{S_i; i = 1, \dots, N\}$, of the input vector space into the sets $S_i = \{x: q(x) = b_i\}$ of input vectors mapping into the i th reproduction vector (or codeword) [1]. Such quantizers are also called block quantizers, vector quantizers, and block source codes [1].

Here the input vectors x can be any kind of customer RFM values. Here it is assumed that the distortion caused by reproducing an input vector x by a reproduction vector i is given by a nonnegative distortion measure $d(x, \hat{x})$. Many such distortion measures have been proposed in the literature. The most common for reasons of mathematical convenience is the squared error distortion, which has been used in the implementation of the algorithm.

$$d(x, \hat{x}) = \sum_{i=0}^{k-1} |x_i - \hat{x}_i| \quad (\text{eq.1})$$

An N-level quantizer will be said to be optimal (or globally optimal) if it minimizes the expected distortion, that is, is optimal if for all other quantizers q having N reproduction vectors $D(q^*) < D(q)$ [1]. A quantizer is said to be locally optimum if $D(q)$ is only a local minimum, that is, slight changes in q cause an increase in distortion [1]. The goal of block quantizer design is to obtain an optimal quantizer if possible and, if not, to obtain a locally optimal and hopefully “good” quantizer [1]. Several such algorithms have been proposed in the literature for the computer-aided design of locally optimal quantizers [1]. A brief view of the steps of this algorithm is given below [1]:

ALGORITHM VQ :

1. Initialization: Given N = number of levels, a distortion threshold $\epsilon \geq 0$, and an initial N-level reproduction alphabet A_0 , and a distribution F . Set $m = 0$ and $D_{-1} = \infty$.

2. Given $A_m = \{y_i; i = 1, \dots, N\}$, find its minimum distortion partition $P(A_m) = \{S_i; i = 1 \dots N\}$: $x \in S_i$ if $d(x, y_i) \leq d(x, y_j)$ for all j . Compute the resulting average distortion, $D_m = D(\{A_m, P(A_m)\}) = E \min_{E_{A_m}} d(X, y)$.
3. If $(D_{m-1} - D_m)/D_m < \epsilon$, halt with A , and $P(A_m)$ describing final quantizer. Otherwise continue.
4. Find the optimal reproduction alphabet $d(P(A_m)) = \{x^*(s_i); i = 1, \dots, N\}$ for $P(A_m)$. set $A_m \triangleq x^*(P(A_m))$. Replace m by $m + 1$ and go to 1.

Earlier this algorithm was mainly used for image compression and other related works. But, I found it useful for my clustering approach.

2.3 The explanation of the VQ algorithm used for the study.

This algorithm for the study of customer segregation consisted of the following components. Firstly, I two double arrays 'c' and 'q' were taken to store the code-vectors and the quantizers values. The an initial 'cref' value is taken as the initial reference value for the code-book, which is the average of all the input vectors x . The input vectors are the data obtained from the sample database, which can be any of the RFM values or all of them. An initial 'dref' value is taken as the mean of all the mean squared distortion values of the input vectors and the quantizers. The order of the Vector quantization algorithm has been denoted by 'n'. The threshold for distortion value is denoted by 'e'.

Initially, the reference code-vector is split into 2 new codevectors by the following process, shown by a pseudocode below :

$$\begin{aligned} c[p-1] &= (1.0+e)*cref[p-1]; \\ c[n-p+1] &= (1.0-e)*cref[p-1]; \end{aligned} \tag{eq.2}$$

Assigning of quantizers to the individual vectors is done by the following method :

```

if((sum[s]-c[p])*(sum[s]-c[p])<min)

    min=(sum[s]-c[p])*(sum[s]-c[p]);

    q[s]=c[p];

```

(eq.3)

updaton of code vectors for the next iteration is done as follows :

```

if (q[s]==c[p]){

    sum3=sum3 + sum[s];

    cnte=cnte +1.0;

if(cnte==0.0)

    c[p]=sum3;

else

    c[p]=sum3/cnte;

```

(eq.4)

CHAPTER 3: The Association rules based approach for customer purchase predictions.

3.1 A background of the Association rules based data mining approach.

Association rules are like classification criteria. There are generally the left-hand side of the rule, known as the antecedent and the right hand side of the rule, known as the inference part. Association rules were initially applied to analyze the relationships of product items purchased by customers at retail stores. In data mining, association rules are descriptive patterns of the form $X \Rightarrow Y$, where X and Y are statements regarding the values of attributes of an instance in a database. X is termed the left-hand-side (LHS), and is the conditional part of an association rule. Meanwhile, Y is called the right-hand-side (RHS), and is the consequent part. The most typical application of association rules is market basket analysis, in which the market basket comprises the set of items (namely itemset) purchased by a customer during a single store visit. It is also known as the Shopping cart analysis of the customer purchases.

The process of Association rule mining approach usually finds out a huge number of rules. These rules are then pruned down on the basis of their “Coverage” value, which is defined as the number of instances from the whole set, where the rule predicts correctly, and their accuracy (the same number expressed as the proportion of instances to which the rule correctly applies). Nowadays, what we call coverage is often called as “support” and what we call

accuracy is also called “confidence”. We are only interested in association rules with high coverage values.

The distinction between LHS and RHS of the association rules seek combinations of attribute – value pairs that have predefines minimum coverage. These are called item-sets. An attribute value pair is called an item. It typically comes from the process of “Market basket analysis” where the store manager analyzes the different items purchased by the customer in a single purchase, and tries to find out association rules among them.

3.2 A Description of the Association rules mining model proposed here.

This study involves the Apriori algorithm, used to detect rules and prune them according to their coverage. ‘Apriori’ is the most basic algorithm for learning association rules. Apriori is designed to operate on databases containing various kinds of transactions (for example, collections of items bought by customers, or details of a website surfing). As is common in association rule mining, given a set of item-sets (for instance, sets of retail transactions, each listing individual items purchased, in this study), the algorithm attempts to find subsets which are common to at least a minimum number K of the itemsets [8]. Apriori uses a "bottom up" approach for its execution, where the required subsets are extended one item at a time (a step known as candidate generation), and such candidates are tested against the data. The algorithm terminates when no further successful items can be added to the existing itemsets. ‘Apriori’, while very basic and historically important, suffers from a number of shortcomings or trade-offs. Candidate generation step generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan of the database). And with increase in the number of itemsets, the computational over head also increases.

It has been considered here, the application of the Apriori algorithm for incremental stages, firstly for a 1-itemset, then 2-itemset and finally for a 3-itemset, each of whom have shown varying results (discussed in the later chapter). The 1 itemset approach applies the formulation of rules based on the criteria as to “if item X purchased”. That is it tries to find out the occurrence of individual items from a number of orders. Then the results are stored , and a pool of rules is obtained which are denoted by “occurrence” here.

For the 2-item set analysis, the market basket is again scrutinized and rules of the form “if item X purchased then item Y purchased” are found out. These are again stored. Now, the coverage of the rules is found out as the ratio of the number of instances where the entire is true for the occurrences and the number of instances where the antecedent is true. Based on this criteria, the rules are pruned again and the qualified rules are stored for further perusal.

For the 3-item set analysis, the market basket analysis is done and rules of the form “if item X and Y purchased then item Z purchased” are found out. These are again stored. Now, the coverage of the rules is found out same as that in 2-item set, as the ratio of the number of instances where the entire is true for the occurrences and the number of instances where the antecedent is true. Based on this criteria, the rules are pruned again and the qualified rules are stored for further perusal.

CHAPTER 4 : Implementation and results.

4.1 Sample dataset and the transformation of data.

In this study sample dataset from the sample database “Nwind.mdb” has been taken, which is a Microsoft Access sample database file, and used the “OrderDeatils” table, which has about 2155 entries, and can be sufficiently large enough for the requirement of the analysis. The data for orderids and the products purchased has been taken, i.e, the first two columns. Then the data has been grouped into market baskets according to the orderIds, i.e., there is 1 market basket for each Order ID. The products mentioned here are given by names. There are about 77 products in all. Then the analysis algorithm assigned integer IDs to each product for keeping the manipulation of the data in the code simple and not confusing. That is, for a certain product, say “tofu” it would have given an id “20” or something like that. The figure on the next page shows a little part of the database table that has been used for doing the analysis work. To keep things simple, only the first 2 columns of the aforementioned table have been utilized, and are sufficient enough to reach a concluding rule set.

OrderID	Product	Unit Pric	Quanti	Discou
10248	Queso Cabrales	\$14.00	12	0%
10248	Singaporean Hokkien Fried Mee	\$9.80	10	0%
10248	Mozzarella di Giovanni	\$34.80	5	0%
10249	Tofu	\$18.60	9	0%
10249	Manjimup Dried Apples	\$42.40	40	0%
10250	Jack's New England Clam Chowder	\$7.70	10	0%
10250	Manjimup Dried Apples	\$42.40	35	15%
10250	Louisiana Fiery Hot Pepper Sauce	\$16.80	15	15%
10251	Gustaf's Knäckebröd	\$16.80	6	5%
10251	Ravioli Angelo	\$15.60	15	5%
10251	Louisiana Fiery Hot Pepper Sauce	\$16.80	20	0%
10252	Sir Rodney's Marmalade	\$64.80	40	5%
10252	Geitost	\$2.00	25	5%
10252	Camembert Pierrot	\$27.20	40	0%
10253	Gorgonzola Telino	\$10.00	20	0%
10253	Chartreuse verte	\$14.40	42	0%
10253	Maxilaku	\$16.00	40	0%
10254	Guaraná Fantástica	\$3.60	15	15%
10254	Pâté chinois	\$19.20	21	15%
10254	Longlife Tofu	\$8.00	21	0%
10255	Chang	\$15.20	20	0%
10255	Pavlova	\$13.90	35	0%
10255	Inlagd Sill	\$15.20	25	0%
10255	Raclette Courdavault	\$44.00	30	0%
10256	Perth Pasties	\$26.20	15	0%
10256	Original Frankfurter grüne Soße	\$10.40	12	0%
10257	Schoggi Schokolade	\$35.10	25	0%
10257	Chartreuse verte	\$14.40	6	0%

Fig 2. The sample dataset for the Association rule mining approach (taken from the table ‘Orderdetails’ “nwind.mdb” MS access database)

The implementation of the customer segmentation used a different table for the VQ approach. It used the Customer column and the freight column of the 'Orders' data table and then used that data to cluster the customers based on the different levels or clusters based on the value of freight. There were about 830 total entries in the Orders table and about 91 customers, each of whom had different freights on different dates. The figure on the next page shows the snap of the table being mentioned and shows some of the entries of the data sets.

So, based on each customer name, the total freight was calculated, and that served as the monetary value for that particular customer. Then that monetary value was taken as the base attribute for the customer segmentation approach. We'll see in the next section as to what segmentations we achieved and how can they be varied and perceived according to the needs of a company.

OrderID	Customer	Freight	Ship Name
10248	Vins et alcools Chevalier	\$32.38	Vins et alcools Chevalier
10249	Toms Spezialitäten	\$11.61	Toms Spezialitäten
10250	Hanari Carnes	\$65.83	Hanari Carnes
10251	Victuailles en stock	\$41.34	Victuailles en stock
10252	Suprêmes délices	\$51.30	Suprêmes délices
10253	Hanari Carnes	\$58.17	Hanari Carnes
10254	Chop-suey Chinese	\$22.98	Chop-suey Chinese
10255	Richter Supermarkt	\$148.33	Richter Supermarkt
10256	Wellington Importadora	\$13.97	Wellington Importadora
10257	HILARIÓN-Abastos	\$81.91	HILARIÓN-Abastos
10258	Ernst Handel	\$140.51	Ernst Handel
10259	Centro comercial Moctezuma	\$3.25	Centro comercial Moctezuma
10260	Ottilies Käseladen	\$55.09	Ottilies Käseladen
10261	Que Delícia	\$3.05	Que Delícia
10262	Rattlesnake Canyon Grocery	\$48.29	Rattlesnake Canyon Grocery
10263	Ernst Handel	\$146.06	Ernst Handel
10264	Folk och få HB	\$3.67	Folk och få HB
10265	Blondel père et fils	\$55.28	Blondel père et fils
10266	Wartian Herkku	\$25.73	Wartian Herkku
10267	Frankenversand	\$208.58	Frankenversand
10268	GROSELLA-Restaurante	\$66.29	GROSELLA-Restaurante
10269	White Clover Markets	\$4.56	White Clover Markets
10270	Wartian Herkku	\$136.54	Wartian Herkku
10271	Split Rail Beer & Ale	\$4.54	Split Rail Beer & Ale
10272	Rattlesnake Canyon Grocery	\$98.03	Rattlesnake Canyon Grocery
10273	QUICK-Stop	\$76.07	QUICK-Stop
10274	Vins et alcools Chevalier	\$6.01	Vins et alcools Chevalier

Fig 3. Sample dataset for the VQ based customer clustering approach. (taken from the Orders table of ‘Nwind.mdb’ MS access database)

4.2 Choice of Programming language and environment.

The study has taken Java as the choice of programming language and Netbeans IDE as the programming environment. The nature of the algorithms that has been implemented, is to work upon the datasets, which can be viewed as objects. In fact, each market basket can be viewed as an object. And the application of java for implementing the algorithms makes it easy to handle the data as objects.

Also, the handling and connectivity of java to the database is very simple and effective. My work being entirely dependent on database records as my sole source of input, it is crucial for effectively getting the data from the Database and then transforming it to the desired form. So, it is easier to handle the database manipulation tasks in java, and it can be very modular, that is the database handling portion of the code can be kept separated from the rest of the algorithm part easily.

4.3 Study of the VQ based clustering algorithm. Results found and discussion.

From the simulation of the algorithm, which had the data of 91 customers of a company, the algorithm calculated the no. of customers in a certain price range of expenditure, using the monetary values obtained from the freight values in the table. Then used the LBG algorithm to cluster the customers according to the spending/freight. And there were about 830 purchases/freight made by all of the customers taken together.

These were the findings or observations :

Spending of customers (the data retrieved from the database, shown in Integer for lucidity):

1357 1394 822 493 2755 364 1403 983 448 1017

1353 5605 2134 58 125 724 821 367 1001 194

1259 6205 3 862 327 1678 623 67 558 469

475 322 1559 64 632 319 281 187 274 432

568 79 97 262 63 363 277 813 6683 191

278 232 635 1186 471 913 306 832 268 37

947 1982 175 207 275 793 673 637 306 75

322 53 219 19 9 168 37 72 1087 458

70 202 88 129 225 171 108 87734

Fig. 4 the initial input data for VQ approach

Using the clustering algorithm, I found the following results :

The no. of people spending less than 100 :18

The no. of people spending between 100 and 500 :37

The no. of people spending between 500 and 1000 :18

The no. of people spending between 1000 and 1500 :9

The no. of people spending more than 1500 :8

Fig 5(a)

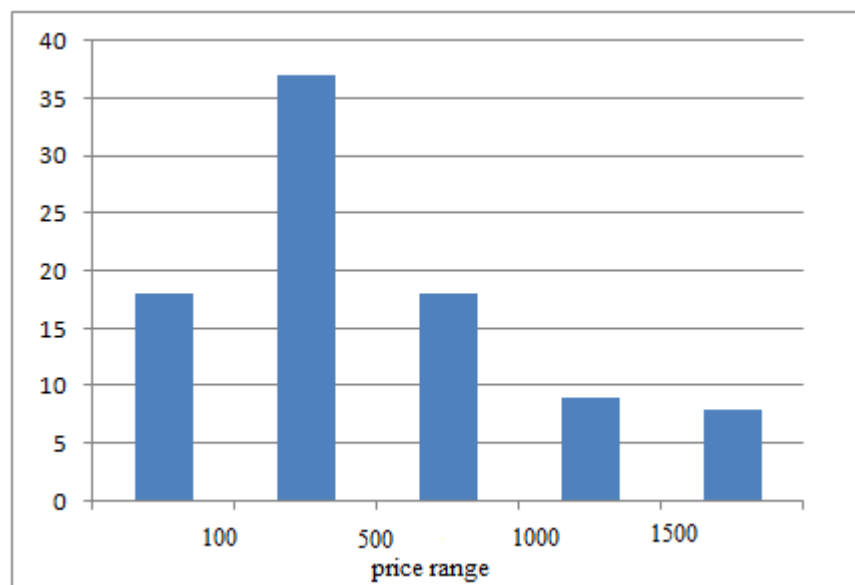


Fig 5(b)

Fig 5(a). The output observed for $e = 0.002$ for the above VQ approach

Fig 5(b). The output shown in a graphical form. X-axis represents the price range and Y-axis represents the number of customers.

4.3.1 Conclusions and inferences from the outputs.

Using the above data, we can form an observation that the number of customers in the range of \$100 to \$500 is large. And the company can predict that laying more stress on providing better services to this range as the customer-base is larger here, would eventually prove more beneficial. Similarly, other metrics can be employed to cluster customers in similar manner, and such conclusions/ inferences can be drawn on the basis of that. This was just a simple implementation to analyze how effectively the VQ based customer clustering approach can segregate customers into groups based on given criteria. The distortion threshold value has been kept 0.002 here for the analysis purposes.

The significance of the distortion measure is that for a certain iteration of the vector quantization loop, if the relative distortion values of two successive distortion values is found to be greater than the threshold value, then the loop is continued again from the previous iteration. The distortion value is a mean squared value of the difference between the particular monetary value and the quantizer associated to it. So, if the difference is too great, then it is safe to repeat the process again and find a new quantizer for the same monetary value and the customer associated with it. This ensures a good clustering of the customers.

4.4 Study of the Association rule based customer behavior mining approach. Results and discussions.

Initially the study applied the Association rule based mining algorithm for 1-itemset, for which I obtained the following output. The occurrence is the number of appearances of the different products in the market baskets. Again, we have only shown a snippet of the output as the output is very lengthy.

```

Occurrence of item 7 : 13.0
Occurrence of item 8 : 5.0
Occurrence of item 9 : 33.0
Occurrence of item 10 : 38.0
Occurrence of item 11 : 14.0
Occurrence of item 12 : 40.0
Occurrence of item 13 : 22.0
Occurrence of item 14 : 6.0
Occurrence of item 15 : 43.0
Occurrence of item 16 : 37.0
Occurrence of item 17 : 27.0
Occurrence of item 18 : 37.0
Occurrence of item 19 : 16.0
Occurrence of item 20 : 39.0
Occurrence of item 21 : 14.0
Occurrence of item 22 : 20.0
Occurrence of item 23 : 51.0
Occurrence of item 24 : 18.0
Occurrence of item 25 : 32.0
Occurrence of item 26 : 9.0
Occurrence of item 27 : 33.0
Occurrence of item 28 : 32.0
Occurrence of item 29 : 32.0
Occurrence of item 30 : 51.0
Occurrence of item 31 : 15.0
Occurrence of item 32 : 32.0
Occurrence of item 33 : 16.0

```

Fig 6. The occurrences for 1- itemset.

Then the study moved on to a 2-itemset approach and devised the rules for 2 – itemsets based on the same datasets. The output had hundreds of rules which then had to be pruned down on the basis of the “coverage values”. Initially I took coverage value as 0.2, for which the following output was shown. :


```
if 2 then 55
if 2 then 59
if 2 then 60
if 2 then 61
if 2 then 68
if 2 then 70
if 4 then 6
if 4 then 7
if 4 then 8
if 4 then 24
if 4 then 39
if 4 then 43
if 4 then 55
if 4 then 58
if 4 then 60
if 4 then 62
if 4 then 73
if 4 then 75
if 4 then 77
if 5 then 29
if 7 then 13
if 7 then 43
if 7 then 46
if 7 then 55
if 7 then 56
if 7 then 60
if 7 then 72
if 8 then 1
if 8 then 2
```

Fig 7. The final rules obtained after pruning with coverage = 0.2.

Next is the pruning results with coverage = 0.5 : figure shown on next page :

```
The final list of rules :  
if 8 then 4  
if 8 then 30  
if 8 then 60  
if 8 then 75  
if 14 then 6  
if 14 then 19  
if 14 then 21  
if 21 then 61  
if 36 then 2  
if 36 then 16  
if 36 then 18  
if 36 then 24  
if 36 then 52  
if 36 then 59  
if 47 then 40  
if 47 then 51  
if 47 then 56  
if 47 then 60  
if 65 then 41  
if 65 then 56
```

Fig 8. The final rules obtained after pruning with coverage = 0.5.

This time we see that there are a lot fewer rules and these rules have a very good chance of turning true because of their high coverage values. These rules can be easily used by the company to draw convenient conclusions for the buying trends of the customers.

Now I tried for the threshold coverage value at 0.7. as we shall see in the figure that follows, a lot fewer rules are obtained than even the previous case, and these rules are even more trustworthy because of their high coverage value of 0.7. figure on next page :

```
if 77 then 32
if 77 then 34
if 77 then 35
if 77 then 38
if 77 then 39
if 77 then 40
if 77 then 41
if 77 then 43
if 77 then 44
if 77 then 46
if 77 then 49
if 77 then 51
if 77 then 52
if 77 then 53
if 77 then 54
if 77 then 55
if 77 then 56
if 77 then 57
if 77 then 59
if 77 then 60
if 77 then 61
if 77 then 62
if 77 then 64
if 77 then 65
if 77 then 66
if 77 then 68
if 77 then 70
```

Fig 9. Rule set with coverage = 0.7.

After that , I went on to analyze the Association rule algorithm for 3-itemset approach. Here, I faced a problem of high execution time of the algorithm due to the enormous amount of iterations being performed in the 3 – nested loops. Then, the loop was modified from a pre-conditioned loop to a post- condition loop which reduced the over head a little, but, it still faces a problem of computational delay due to the sheer number of iterations. The observations recorded are shown on the figure :

```

if 39 and 75, then 46
if 39 and 75, then 52
if 39 and 75, then 55
if 39 and 75, then 60
if 39 and 75, then 64
if 39 and 75, then 66
if 39 and 75, then 73
if 39 and 75, then 77
if 39 and 77, then 2
if 39 and 77, then 3
if 39 and 77, then 4
if 39 and 77, then 6
if 39 and 77, then 7
if 39 and 77, then 8
if 39 and 77, then 10
if 39 and 77, then 12
if 39 and 77, then 13
if 39 and 77, then 14
if 39 and 77, then 16
if 39 and 77, then 20
if 39 and 77, then 23
if 39 and 77, then 27
if 39 and 77, then 32
if 39 and 77, then 41
if 39 and 77, then 46
if 39 and 77, then 52
if 39 and 77, then 55
if 39 and 77, then 60
if 39 and 77, then 64
if 39 and 77, then 66
if 39 and 77, then 73
if 39 and 77, then 75
if 40 and 1, then 8
if 40 and 1, then 21
if 40 and 1, then 28
if 40 and 1, then 30
if 40 and 1, then 36
if 40 and 1, then 52
if 40 and 1, then 53
if 40 and 7, then 17
if 40 and 7, then 33
if 40 and 7, then 72
if 40 and 8, then 1
if 40 and 8, then 21
if 40 and 8, then 30
if 40 and 8, then 44
if 40 and 10, then 31
if 40 and 10, then 33
if 40 and 10, then 76
if 40 and 11, then 57

```

Fig 10. The snippet of rule set for 3-itemset obtained for coverage = 0.5

It took about 96 mins to complete the execution of the whole program and the final output had to be stored on a separate text file for future references. For the 3-itemset purpose I took the value of the coverage as 0.5.

CHAPTER 5 : CONCLUSIONS

5.1 Conclusions.

On implementing the approaches in sections 4.3 and 4.4, it was observed that the two approaches show varied resulting data that can be interpreted in different ways:

- VQ approach can be basically used to segment customers, according to any of the RFM values, or all of them together. It needs the initial vectors as its input for it to start creating the clusters.
- The change of the level N of the algorithm results in more number of segmentations, and is to be adjusted according to the requirements of the clustering approach.
- The threshold value ‘ ϵ ’ strictly keeps a check on the difference between the customer’s monetary value (or for that matter, any other value required) and its respective quantizing value found by the algorithm.
- In terms of predictions, the clusters obtained can show the different segments of customers and the more populated segments can be targeted specifically.
- The customer purchase patterns approach, using the association rules mining technique, is an effective way of extracting the rules from the raw data and inferring the buying patterns among them.
- The implementation shows that increasing the “coverage” values results in better pruning of rules, and a more trustworthy rule set.
- Also, the increase from a 1-item to a 2-itemset and then onto a 3-itemset, results in a 10-folds increase in the computation times of the algorithms. The changing of loop-nesting from a pre-conditioned one to the post-condition one does reduce the execution time a bit but not very significantly.

- From the association rules with sufficient coverage, we can predict which products the customer tends to buy along with the purchase of particular products.

5.2 Future work.

The work done so far leaves ample amount of space for future improvements and comparisons with other algorithms. The Association rule algorithm implemented here takes up a lot of execution time. So, the optimization of the algorithm can be done to ensure a better performing algorithm. The VQ algorithm implemented here doesn't take into account the other aspects like Recency and Frequency values. Either, these values can be taken into account 1 at a time or, all three at a time using weight factors for the three parameters. Of course, here a simple database was used, so there isn't much space for taking the frequency values into account. But it can be extended to a larger and more comprehensive database to analyze the aforementioned values.

REFERENCES

- [1] Yoseph Linde, Andres Buzo, Robert M. Gray : An Algorithm for Vector Quantizer Design, *IEEE Transactions on communications*, vol. com-28, no. 1, (january 1980), pp. 84-86.

- [2] Danuta Zakrzewska, Jan Murlewski : Clustering Algorithms for Bank Customer Segmentation, *5th International Conference on Intelligent Systems Design and Applications*, (2005),pp 1-2.

- [3] Abdullah Al-Mudimigh, Farrukh Saleem, Zahid Ullah Department of Information System: Efficient implementation of data mining: improve customer's behavior, *2009 IEEE* ,(2009),pp.7-10.

- [4] Sung Ho Ha , Sang Chan Park, Sung Min Bae : Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case, *Computers & Industrial Engineering* 43 (2002) 801–820, (2002),pp.801-806.

- [5] Euiho Suh, Seungjae Lim, Hyunseok Hwang, Suyeon Kim : A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study, *Expert Systems with Applications* 27 ,(2004), pp. 245-250.

[6] Mu-Chen Chen , Hsu-Hwa Chang, Ai-Lun Chiu : Mining changes in customer behavior in retail marketing, *Expert Systems with Applications* 28 ,(2005), pp. 773-776.

[7] Sriram Thirumalai, Kingshuk K. Sinha : Customer satisfaction with order fulfillment in retail supply chains: implications of product type in electronic B2C transactions, *Journal of Operations Management* 23 ,(2005), pp. 291-296.

[8] Ian H. Witten & Eibe Frank : Data Mining : Practical machine learning tools and techniques, *San Francisco, Morgan Kaufmann publishers*, (2005), pp. 112-118,136-139.